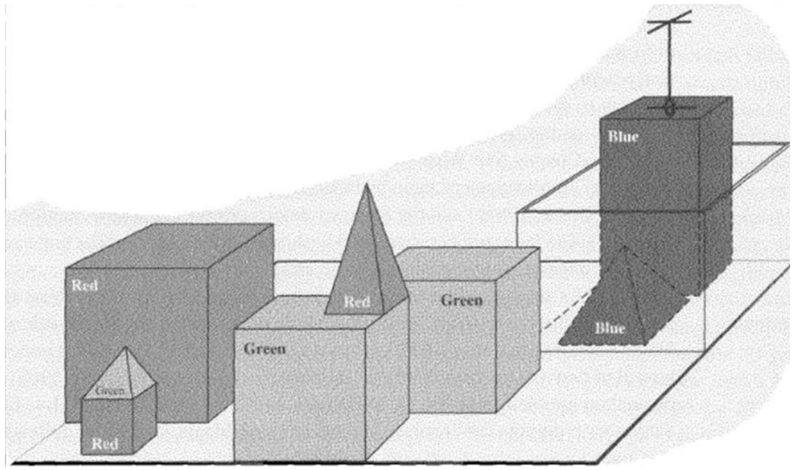
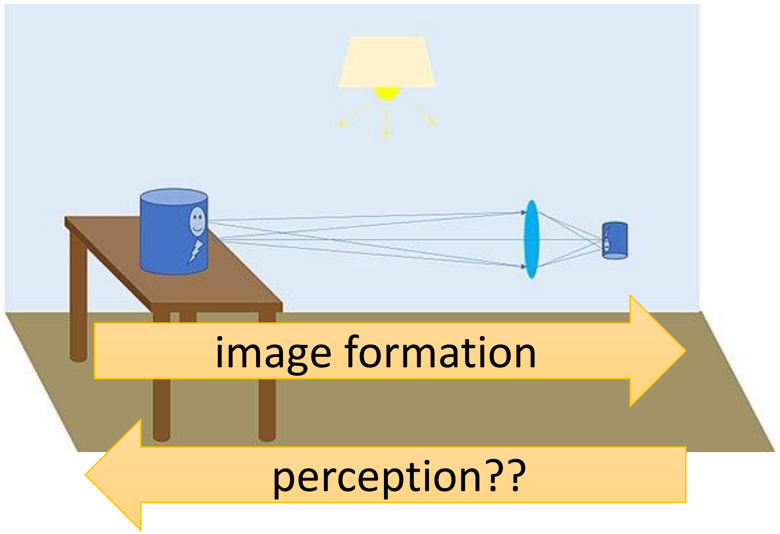
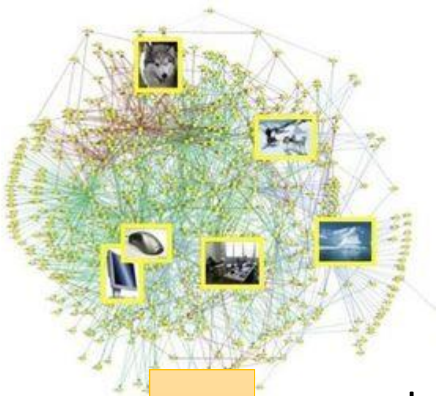
# Embodied Implicit Scene Understanding

**Sergey Levine**

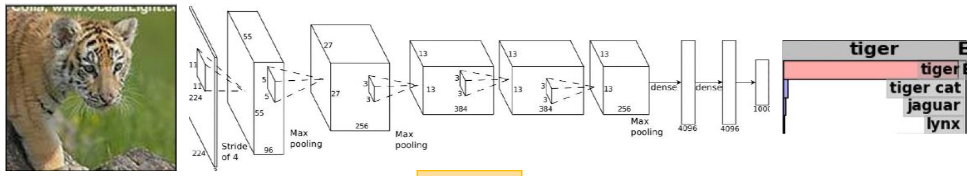
**UC Berkeley**



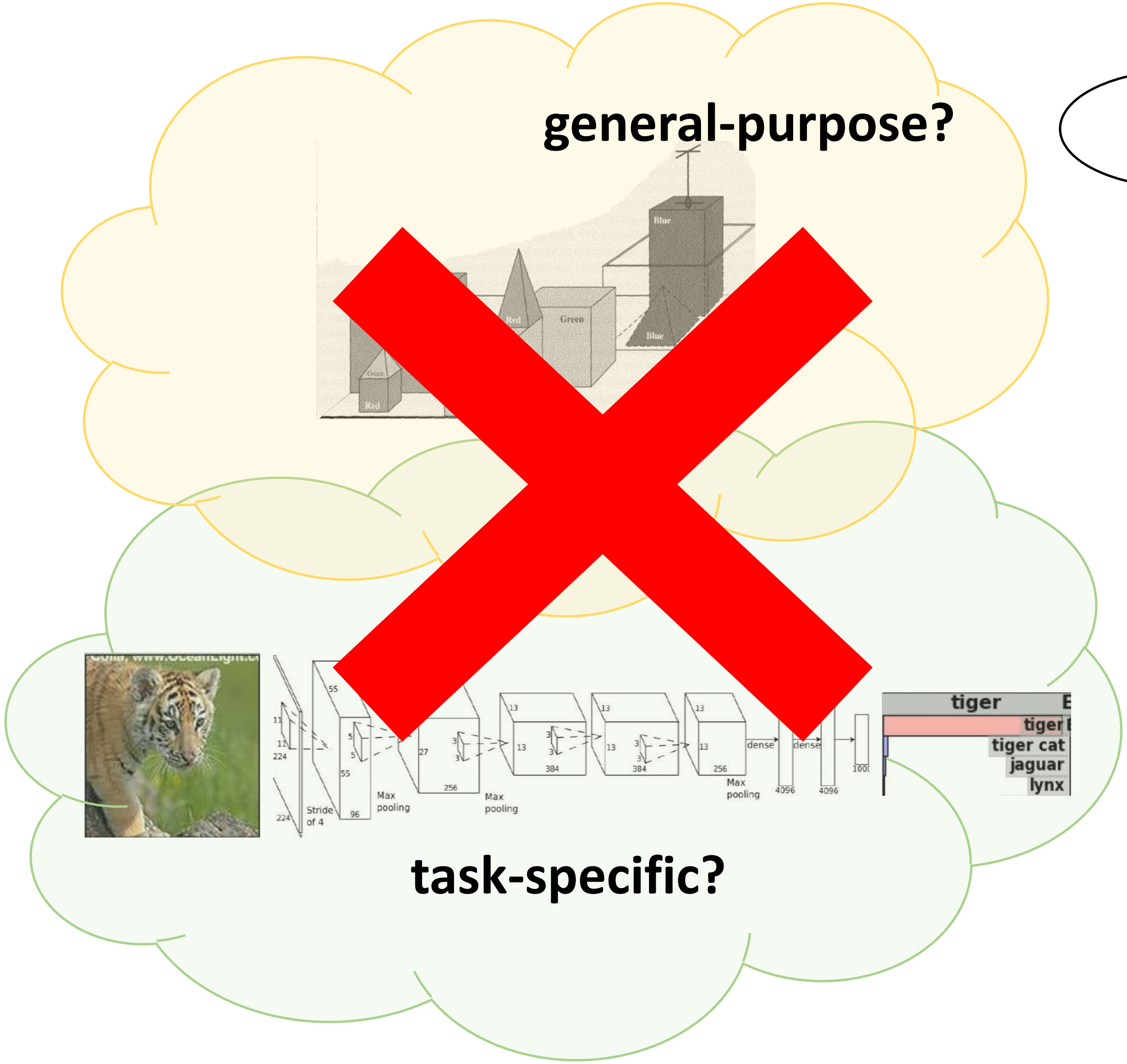



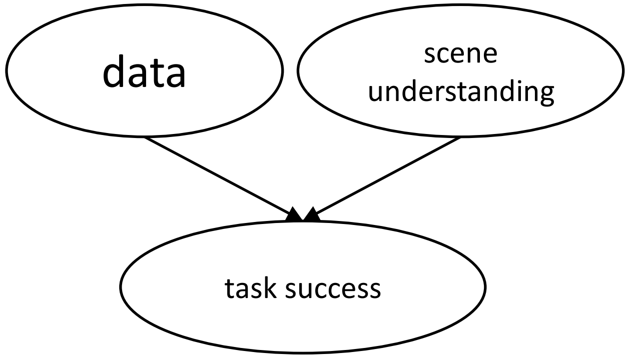
end-to-end training



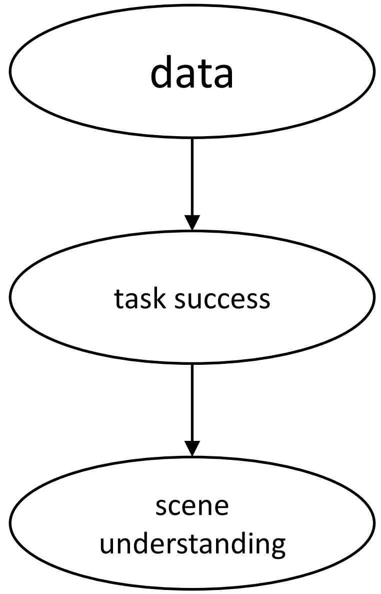




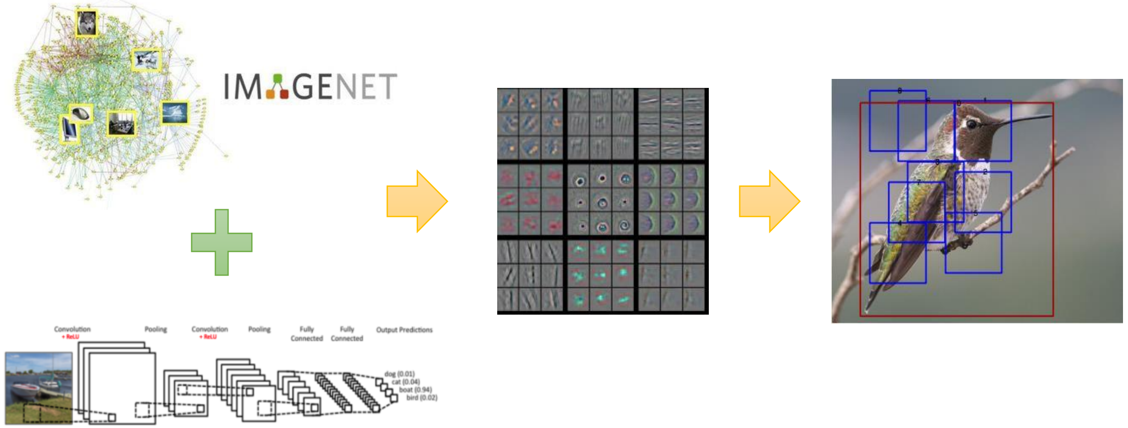

"classic" view




"epiphenomenon view"




end-to-end systems learn **general** concepts!



# Understanding the world via end-to-end learning?

Our universe:

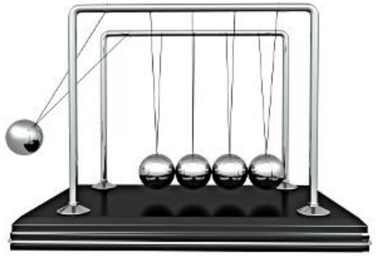
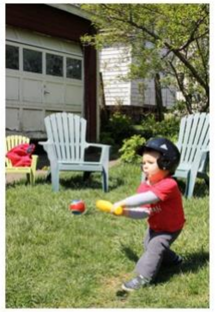




Image captioning "universe":


a group of people standing around a room with remotes
logprob: -9.17



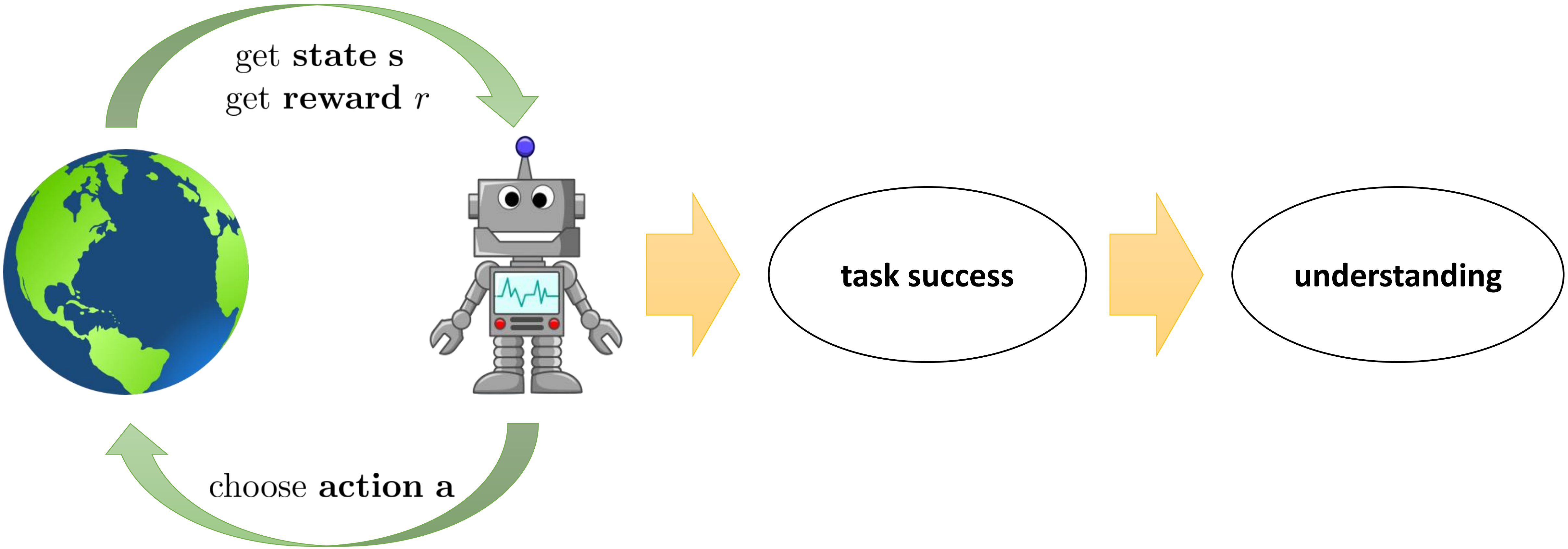
a young boy is holding a baseball bat
logprob: -7.61

a toilet with a seat up in a bathroom
logprob: -13.44

a woman holding a teddy bear in front of a mirror
logprob: -9.65

Karpathy & Li, 2016

# An **embodied learning** recipe for scene understanding?



# Which end-to-end task should we use?

**Model-free algorithms:**
predict future *rewards*

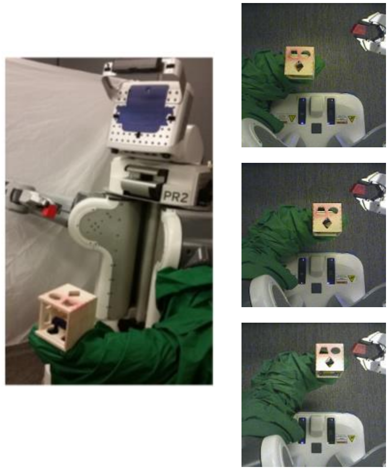
**Model-based algorithms:**
predict future *observations*

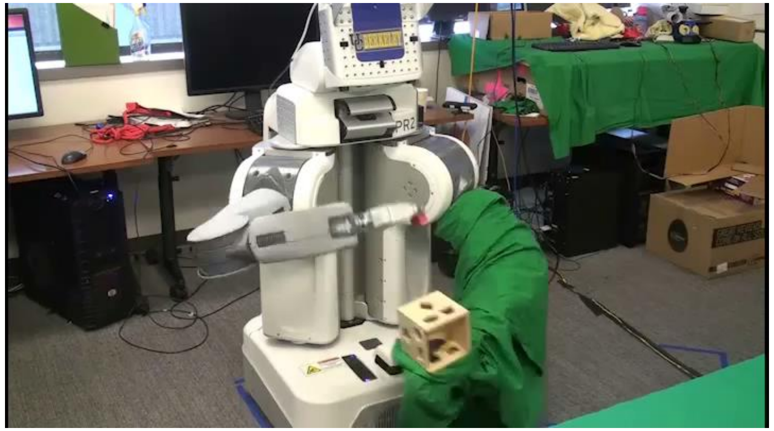
# Which end-to-end task should we use?

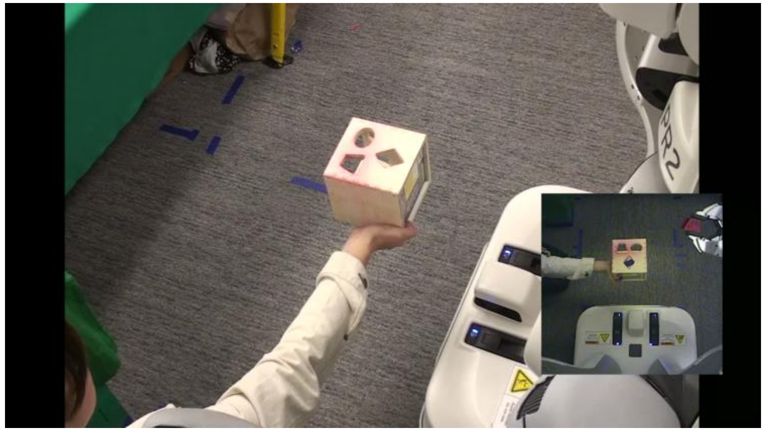
**Model-free algorithms:**
predict future *rewards*

**Model-based algorithms:**
predict future *observations*

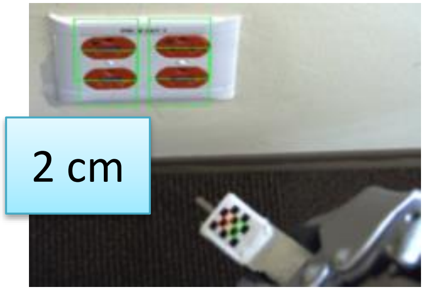
# End-to-end training



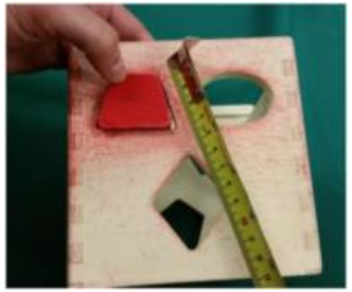




| network architecture | test error (cm) |
|---|---|
| softmax + feature points (**ours**) | **1.30 ± 0.73** |
| softmax + fully connected layer | 2.59 ± 1.19 |
| fully connected layer | 4.75 ± 2.29 |
| max-pooling + fully connected | 3.71 ± 1.73 |



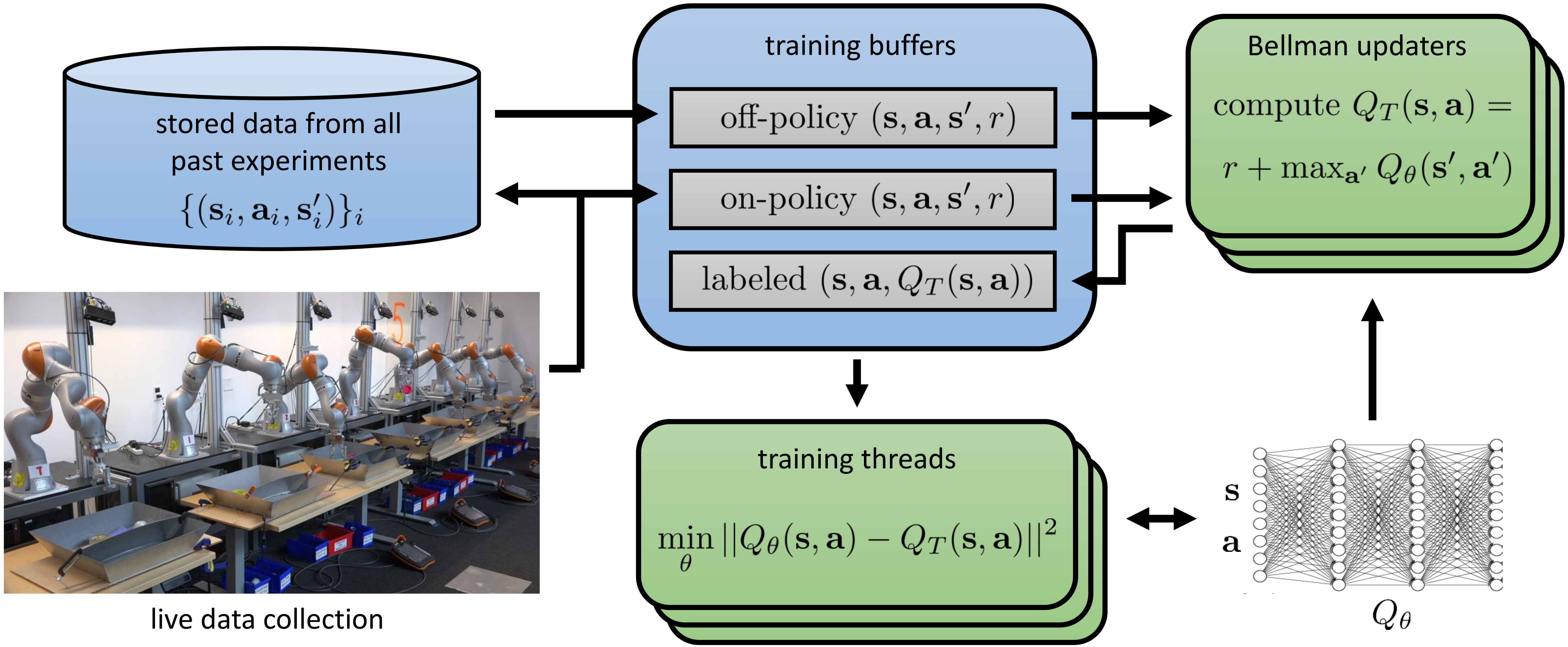

Meeussen et al. (WG)



| shape sorting cube | success rate |
|---|---|
| pose prediction | 0% |
| pose features | 70.4% |
| end-to-end training | **96.3%** |

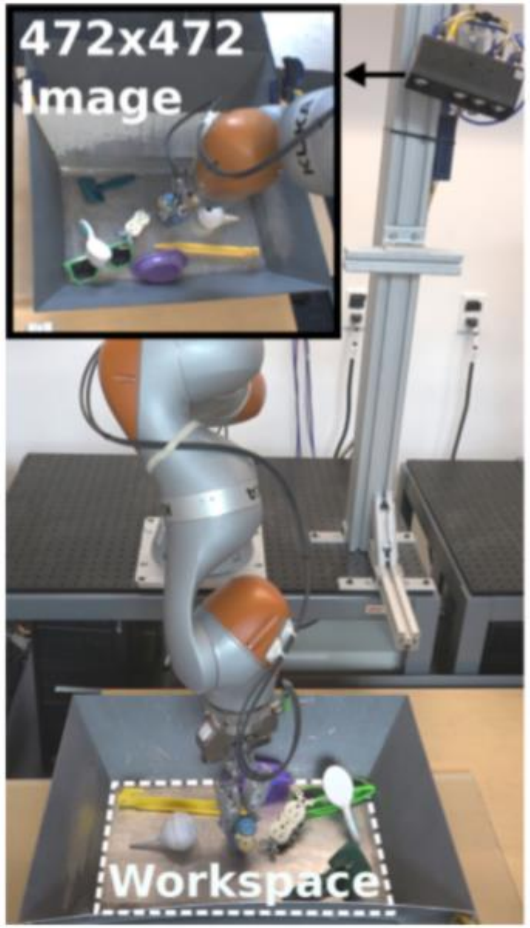
Levine*, Finn*, Darrell, Abbeel. **End-to-End Training of Deep Visuomotor Policies.**

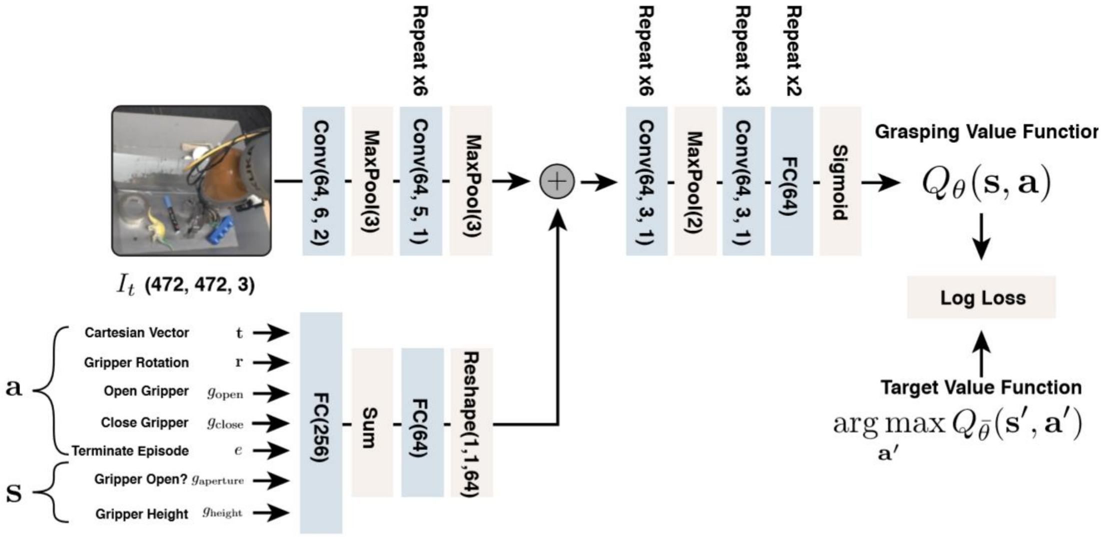
# QT-Opt: robotic RL at scale




Kalashnikov, Irpan, Pastor, Ibarz, Herzong, Jang, Quillen, Holly, Kalakrishnan, Vanhoucke, Levine. **QT-Opt: Scalable Deep Reinforcement Learning of Vision-Based Robotic Manipulation Skills**

# Grasping with QT-Opt




- About 1000 training objects
- About 600k training grasp attempts
- Q-function network with 1.2M parameters
- The only grasp-specific feature is the reward (1 if grasped)

Kalashnikov, Irpan, Pastor, Ibarz, Herzong, Jang, Quillen, Holly, Kalakrishnan, Vanhoucke, Levine. **QT-Opt: Scalable Deep Reinforcement Learning of Vision-Based Robotic Manipulation Skills**

# Emergent grasping strategies

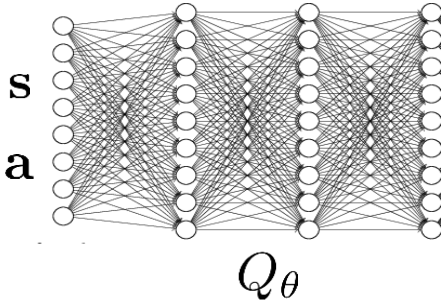


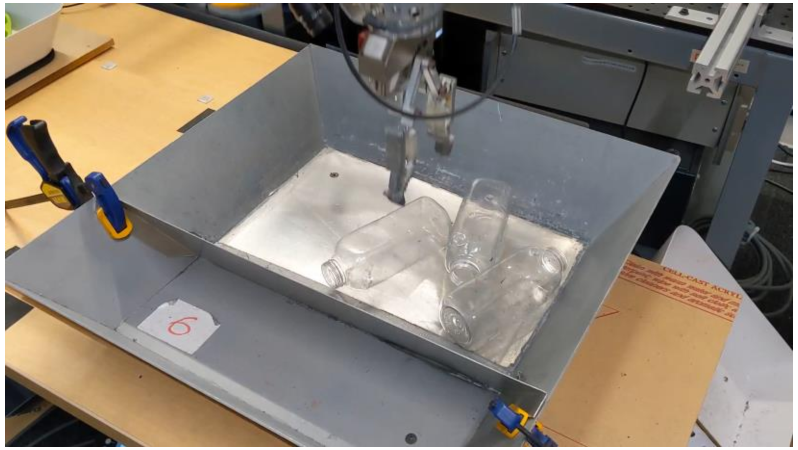
**96%**

# Learning on the job

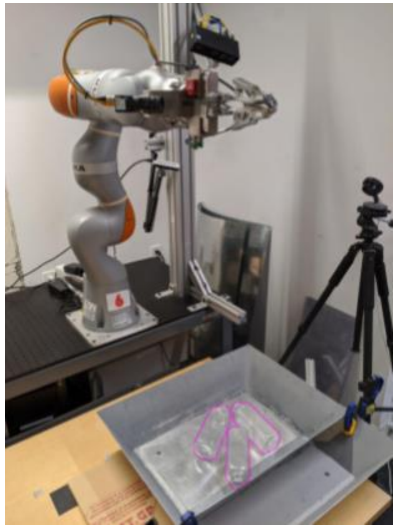
**training**

96% **success rate**



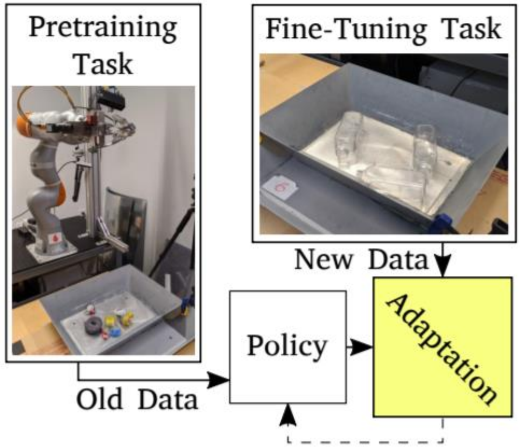
**“on the job”**



49% **success rate**

## just keep training!

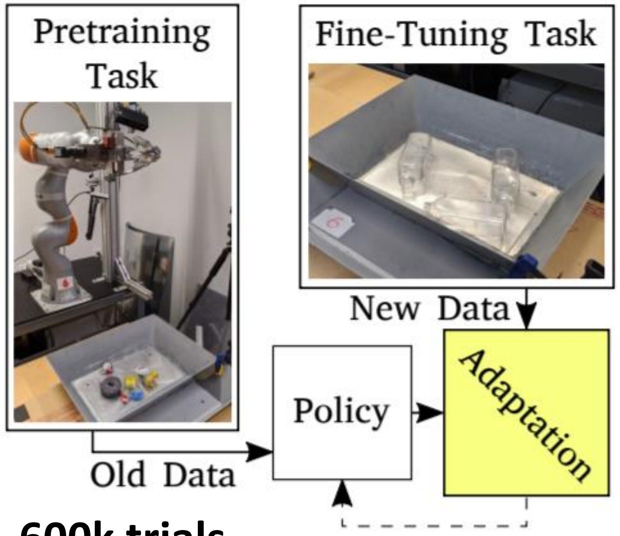
no human effort required






Julian, Levine, Finn, Hausman. **Efficient Policy Adaptation for End-to-End Vision-Based Robotic Manipulation.**

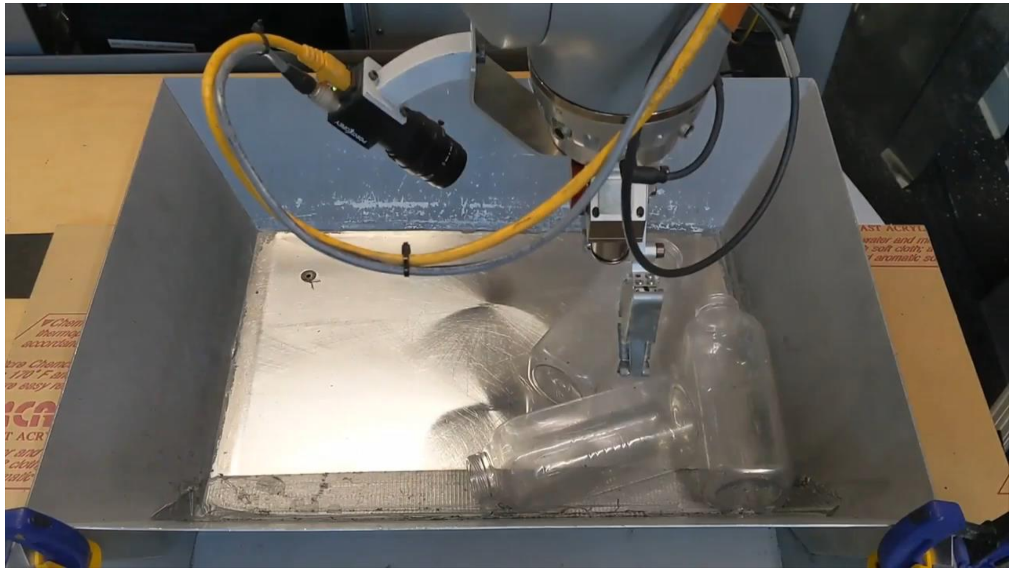
# Learning on the job

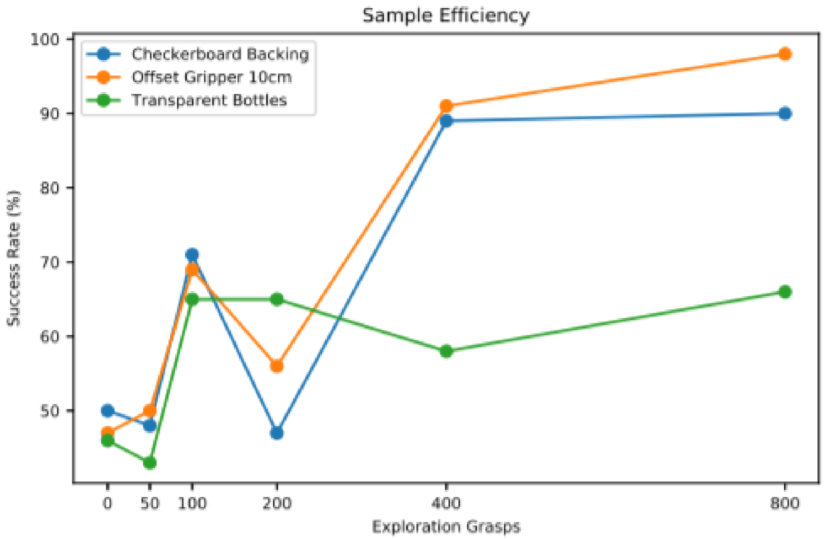


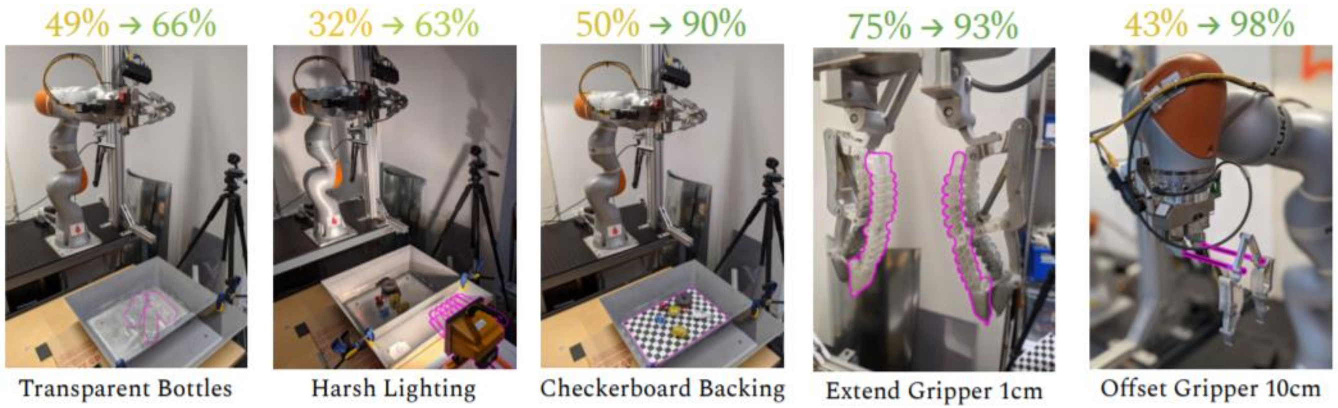

800 trials

about 4 hours

600k trials








Julian, Levine, Finn, Hausman. **Efficient Policy Adaptation for End-to-End Vision-Based Robotic Manipulation.**

# Grasping provides supervision



pre-grasp scene



grasped object



post-grasp scene

Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

# Grasping provides supervision



pre-grasp scene

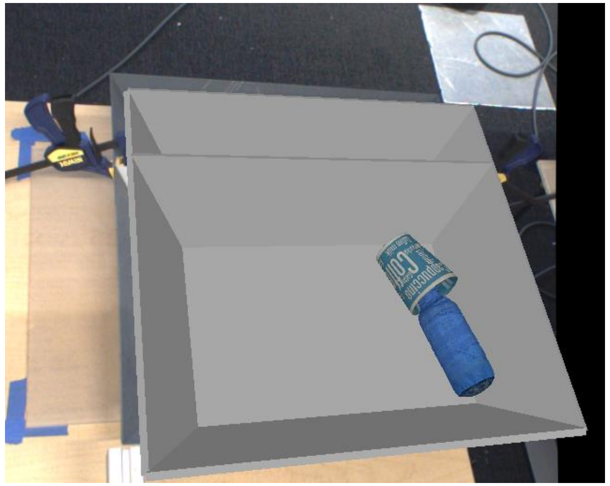
grasped object



post-grasp scene

Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

# Grasping provides supervision



pre-grasp scene

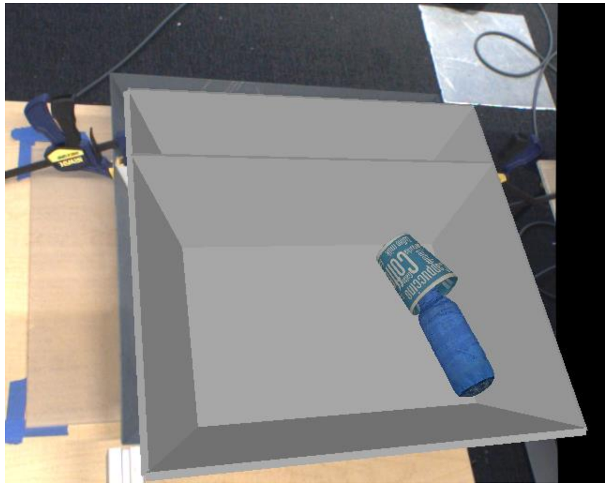
grasped object



post-grasp scene

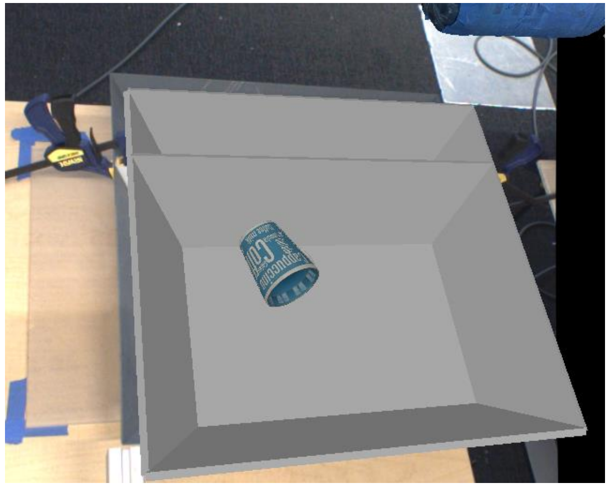
Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

# Grasping provides supervision



pre-grasp scene

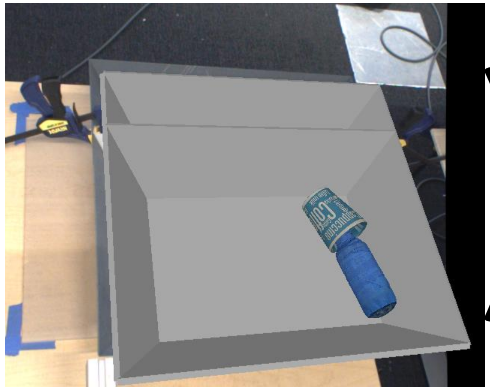
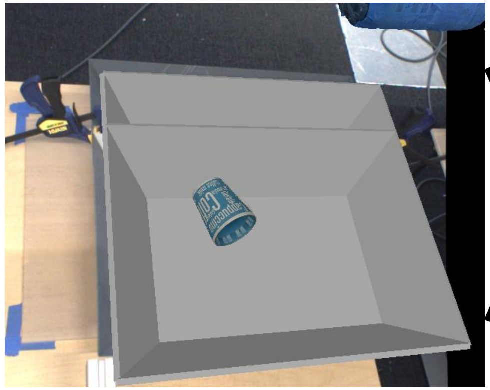
grasped object



post-grasp scene

Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

# Representation learning from grasping

$\varphi($  $) - \varphi($  $) = \varphi($  $)$

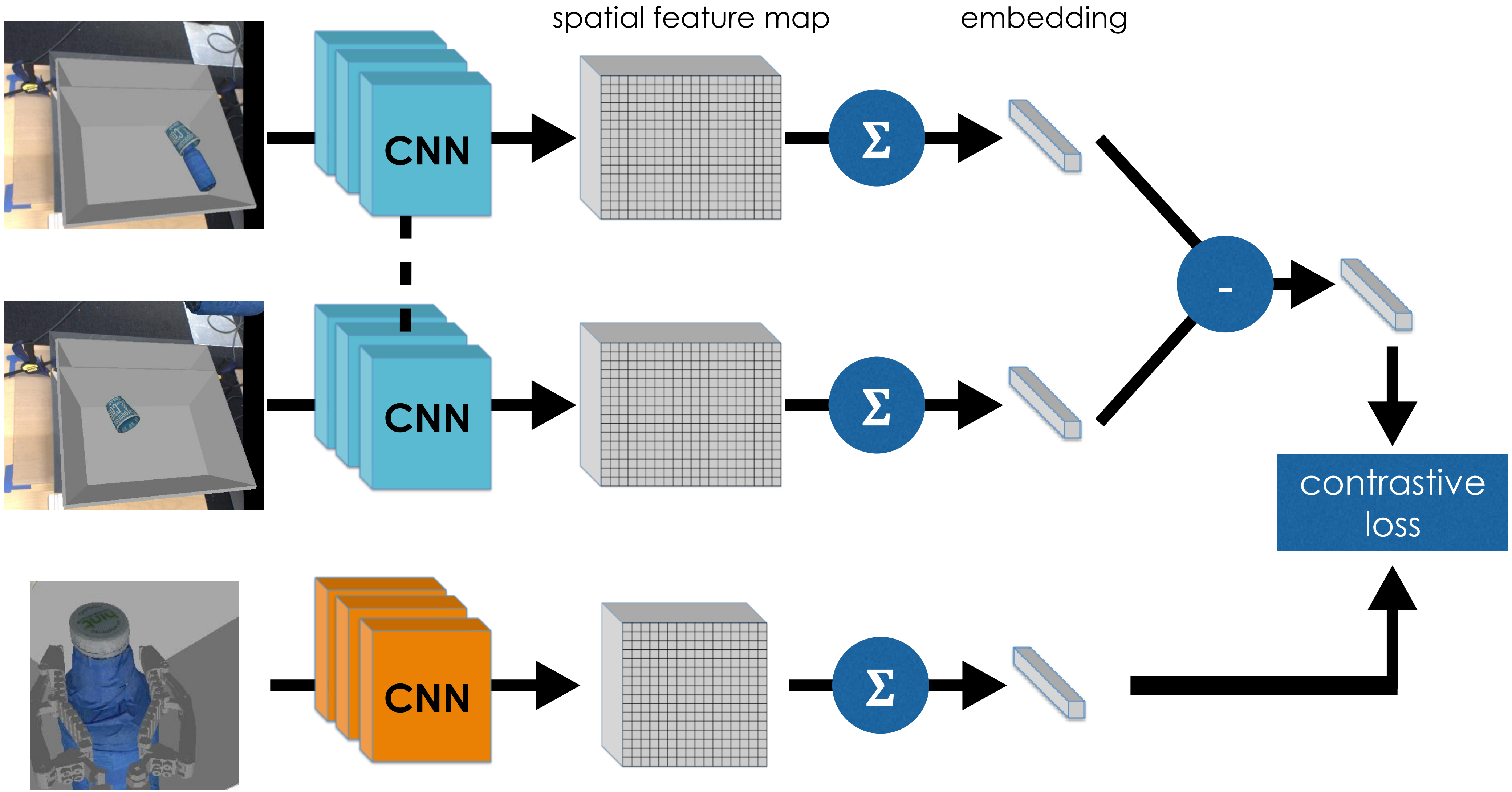
pre-grasp scene

post-grasp scene

grasped object

Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

# Training with contrastive loss




Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

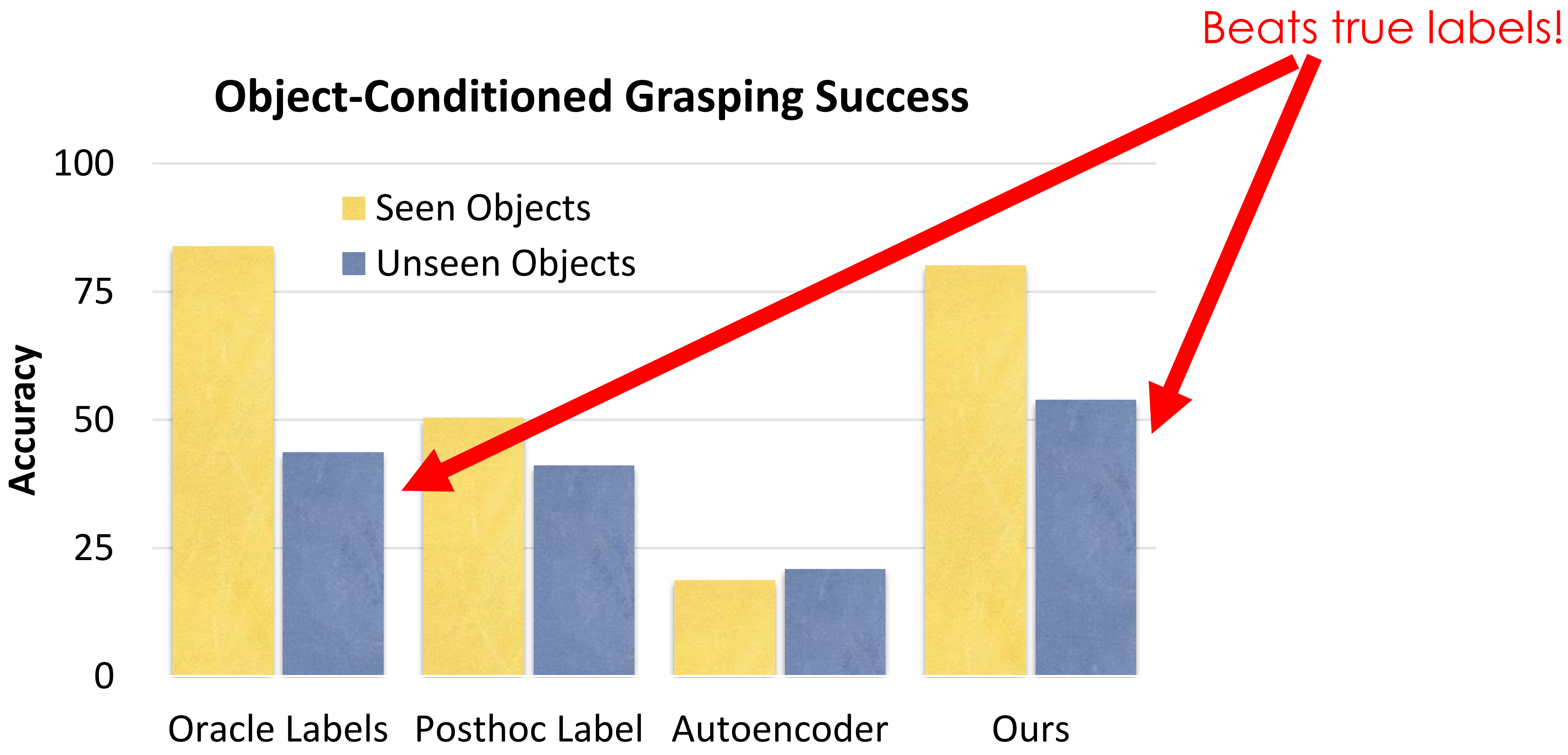
# What can we do with the learned representation?

Object-specific grasping


Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

# What can we do with the learned representation?

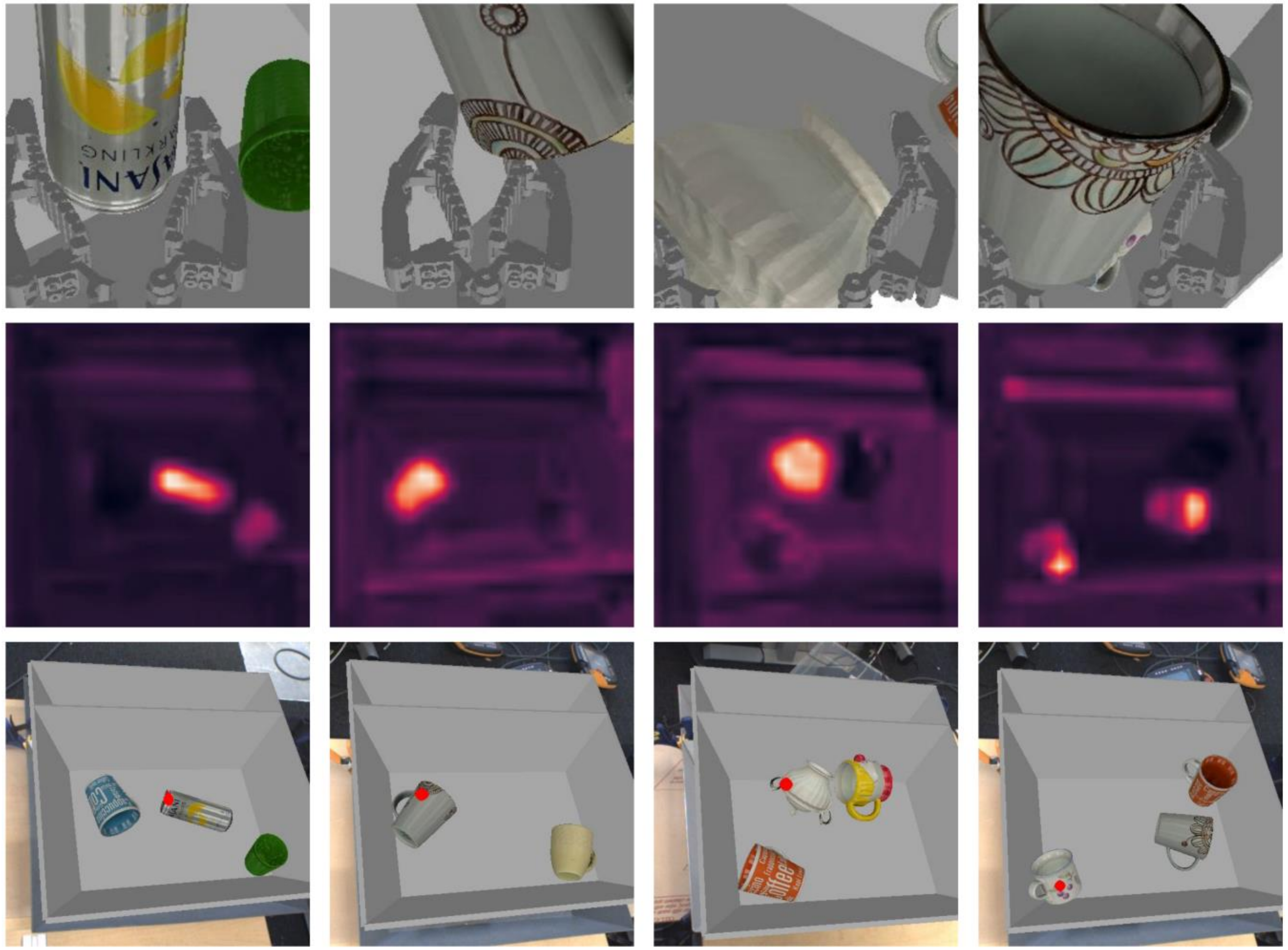
## Object-specific grasping




Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

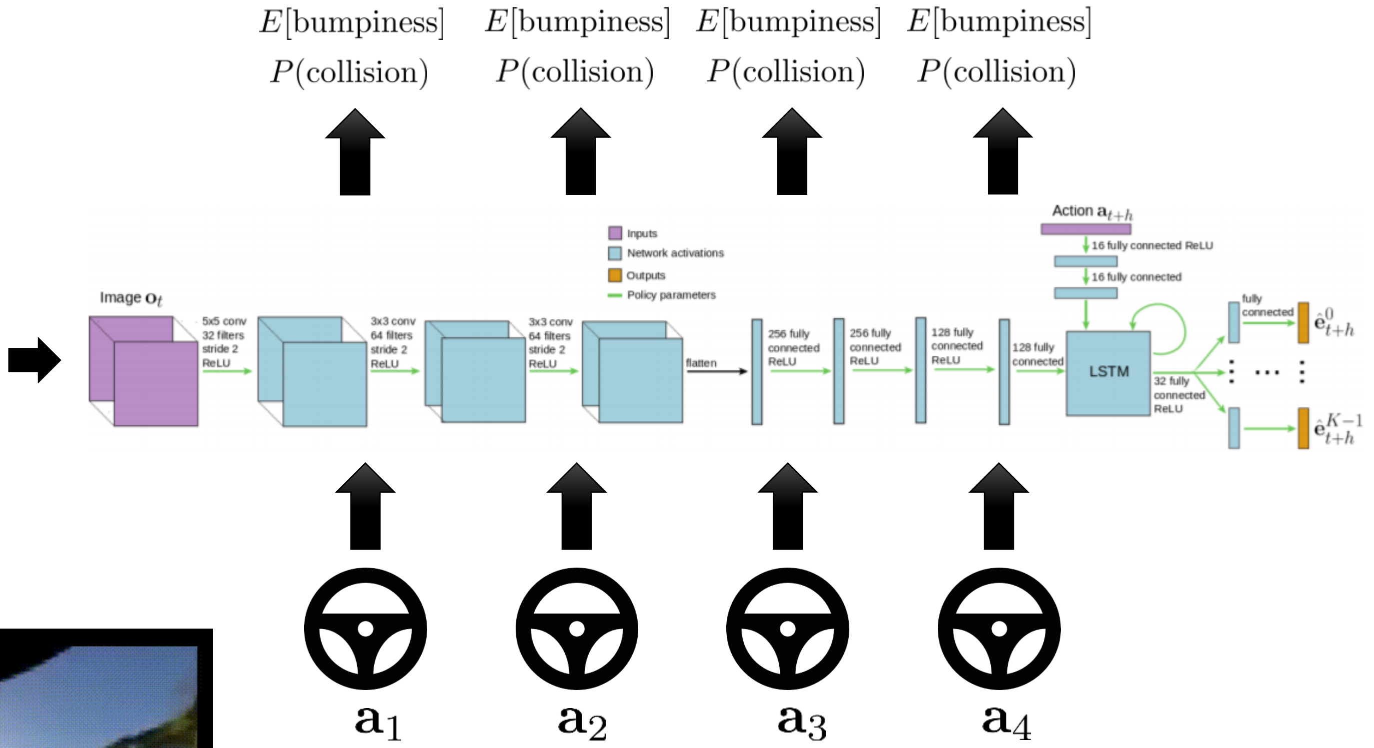
# What can we do with the learned representation?

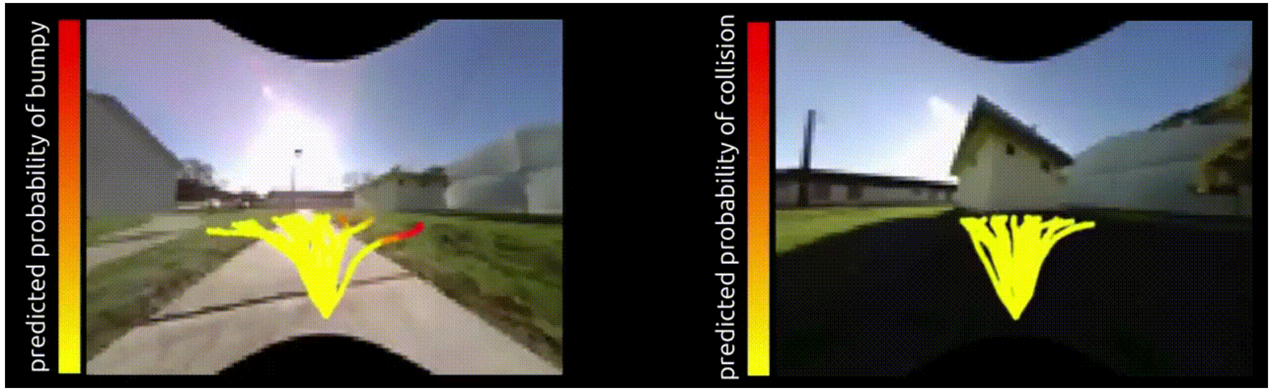
## Fully self-supervised localization



Jang*, Devin*, Vanhoucke, Levine. **Grasp2Vec: Learning Object Representations from Self-Superivsed Grasping.** CoRL 2018.

# Can we learn to understand *open-world* scenes?






Kahn, Abbeel, Levine. **BADGR: An Autonomous Self-Supervised Learning-Based Navigation System.**

# Navigational affordances


**baseline method** **our method**

- Learns a kind of "navigational common sense" from experience
- Some obstacles (e.g., grass) are traversable
- Concrete paths are good for avoiding bumpiness



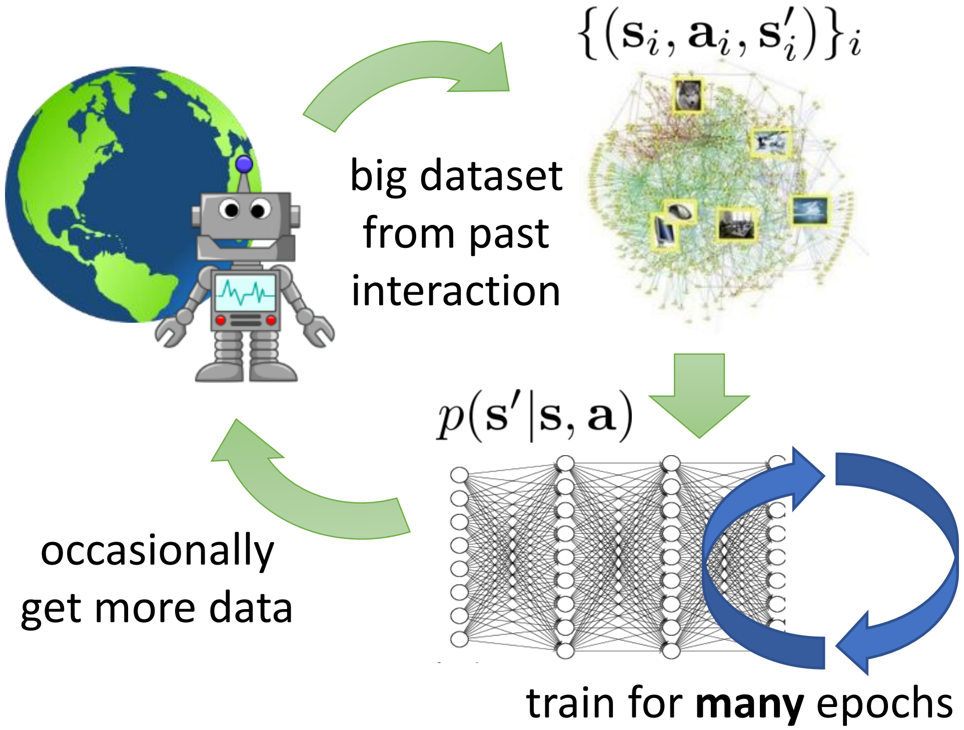
Kahn, Abbeel, Levine. **BADGR: An Autonomous Self-Supervised Learning-Based Navigation System.**

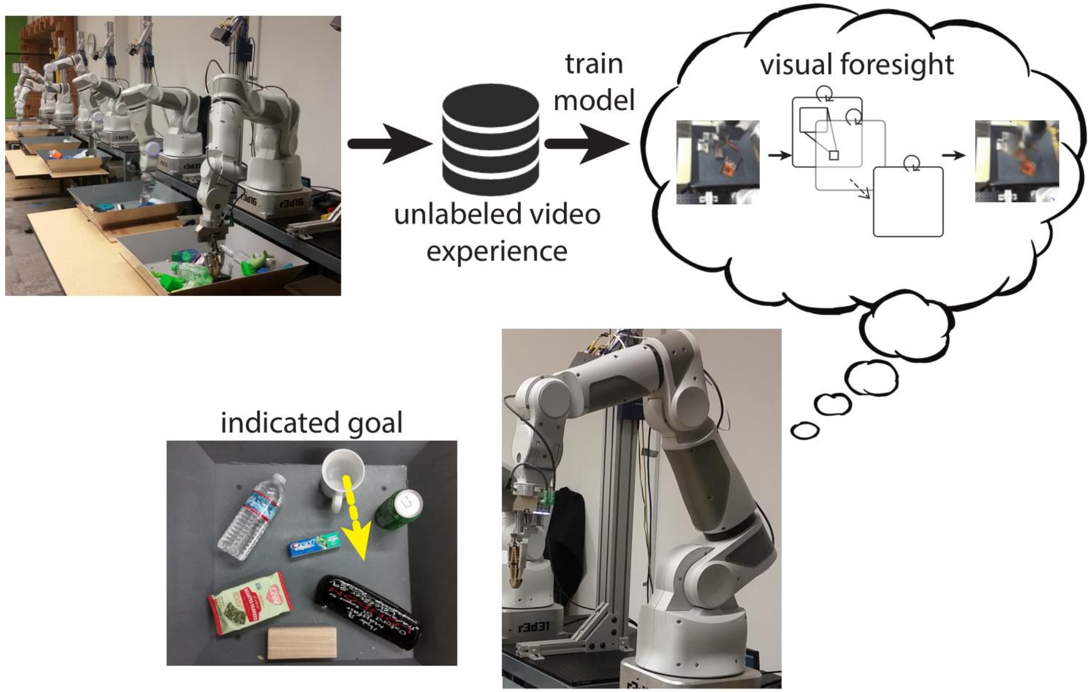
# Which end-to-end task should we use?

**Model-free algorithms:**
predict future *rewards*

**Model-based algorithms:**
predict future *observations*

# Learning to predict the future



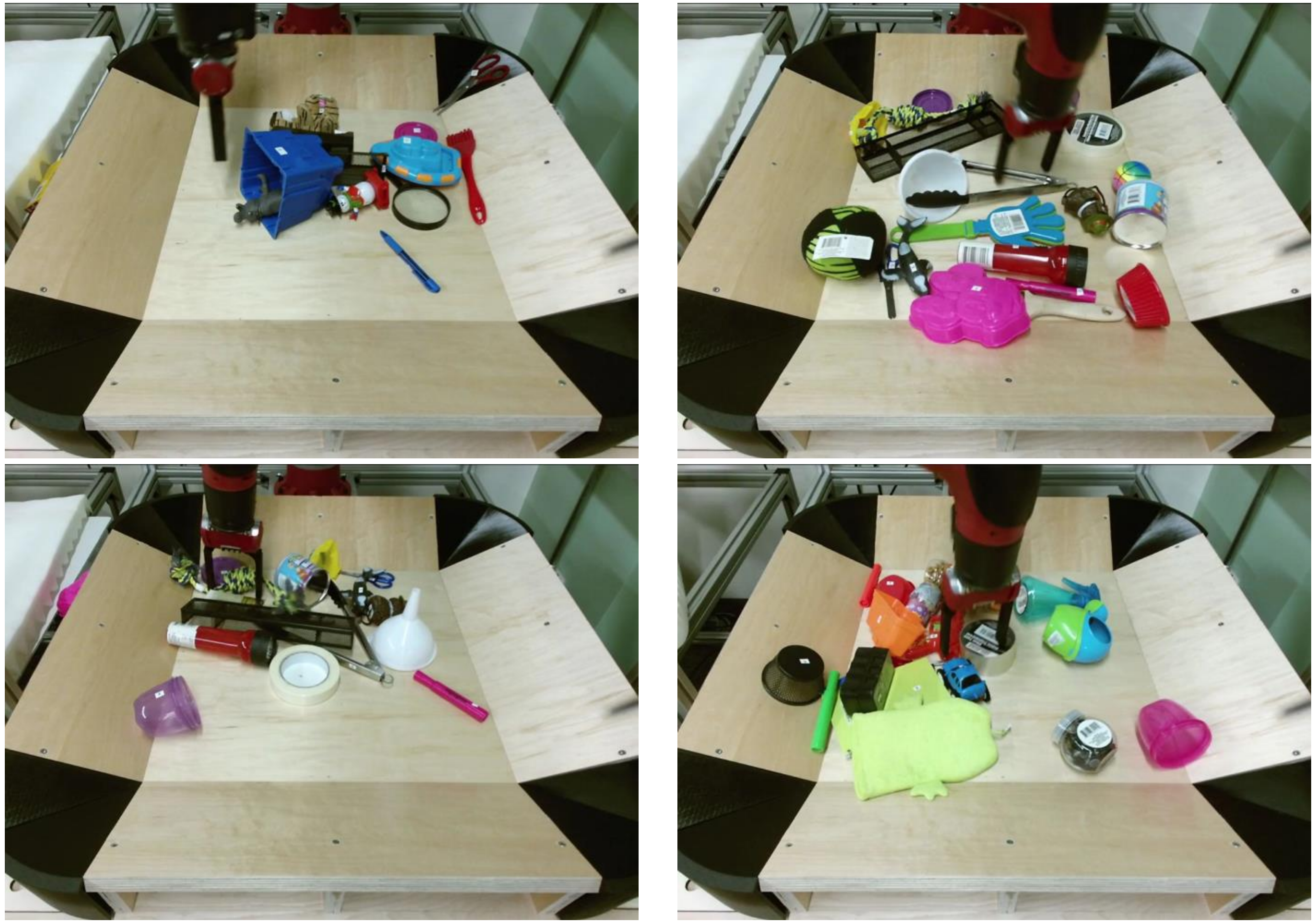




Finn, Levine. **Deep Visual Foresight for Planning Robot Motion.**

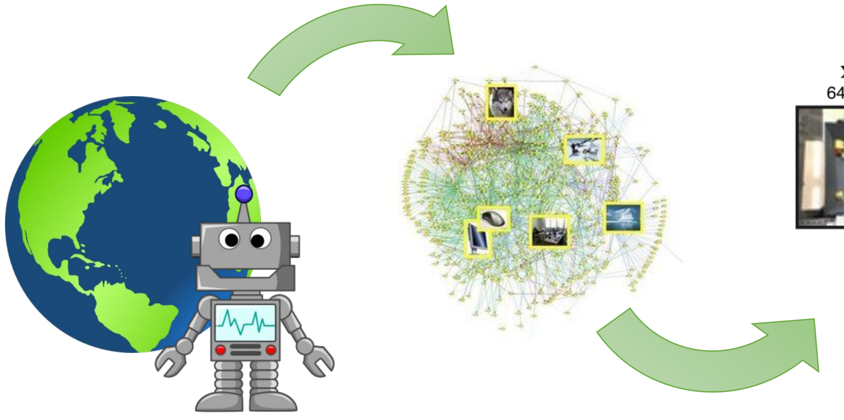
Ebert, Finn, Lee, Levine. **Self-Supervised Visual Planning with Temporal Skip Connections.**

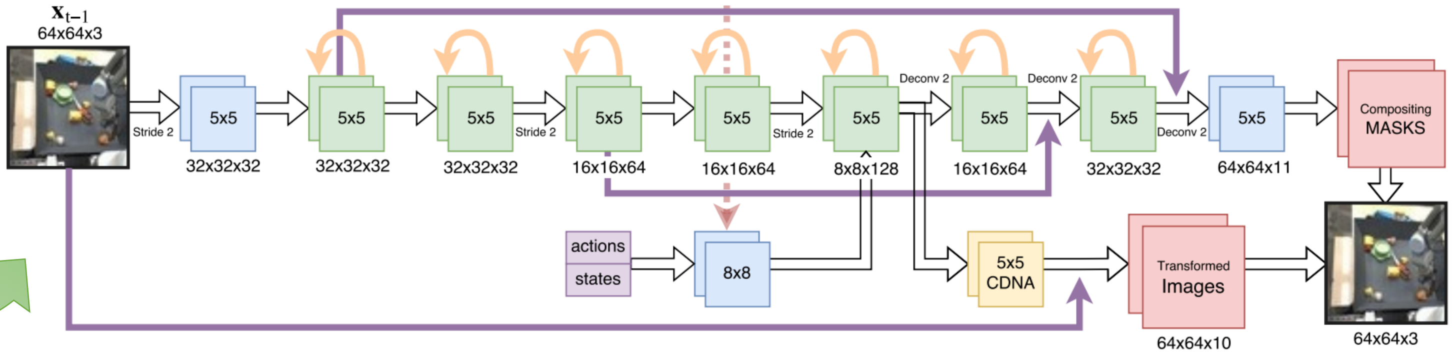
Lee, Zhang, Ebert, Abbeel, Finn, Levine. **Stochastic Adversarial Video Prediction.**

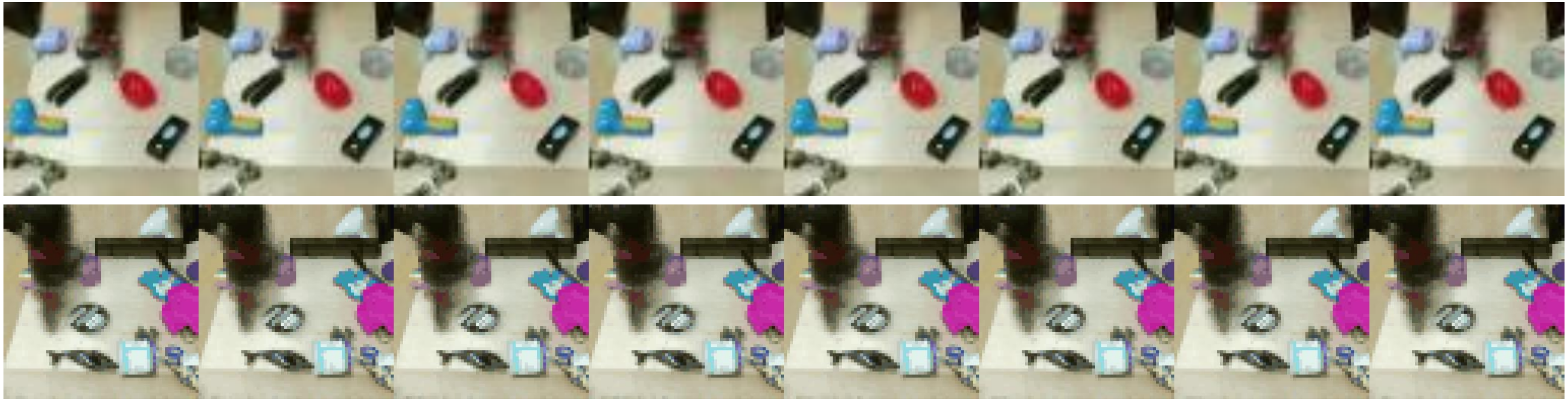
# Collect data by playing with objects



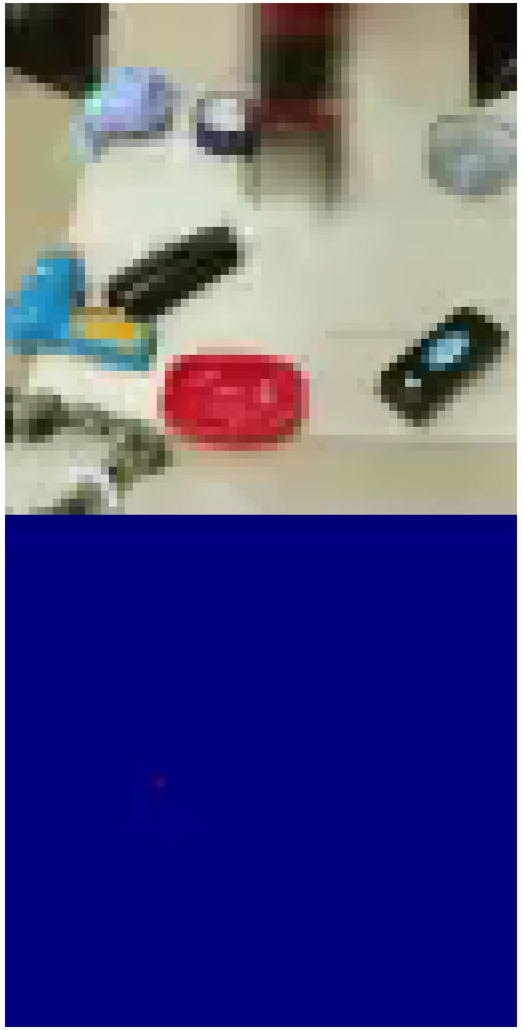
# Learn to predict the future

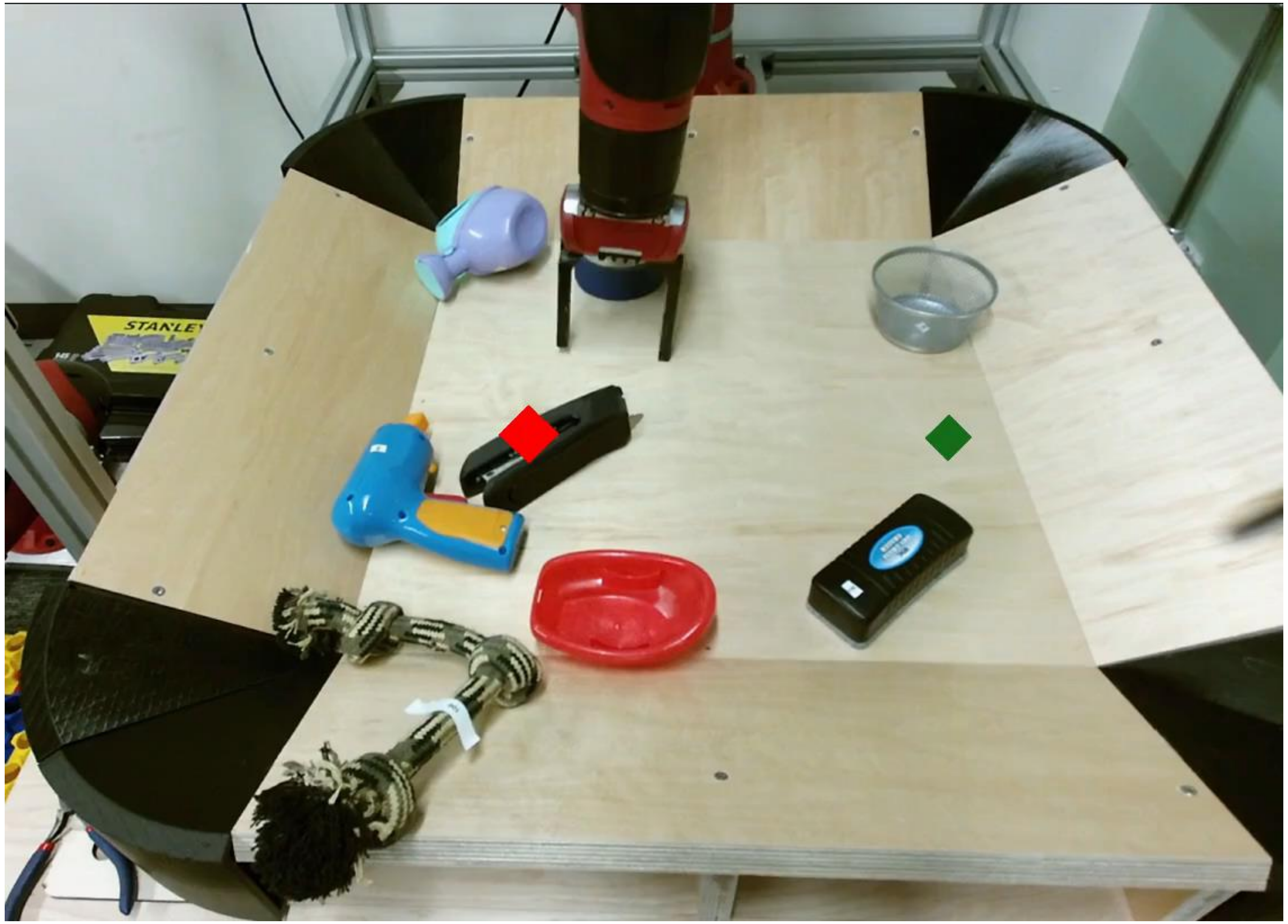




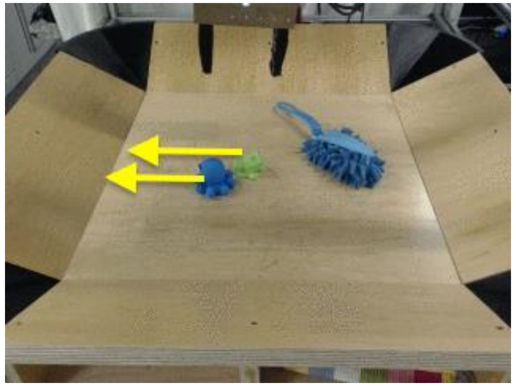
# Example execution

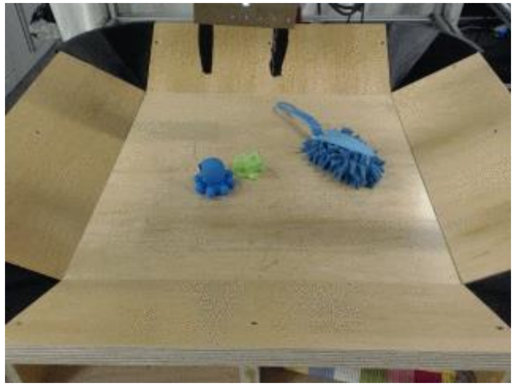


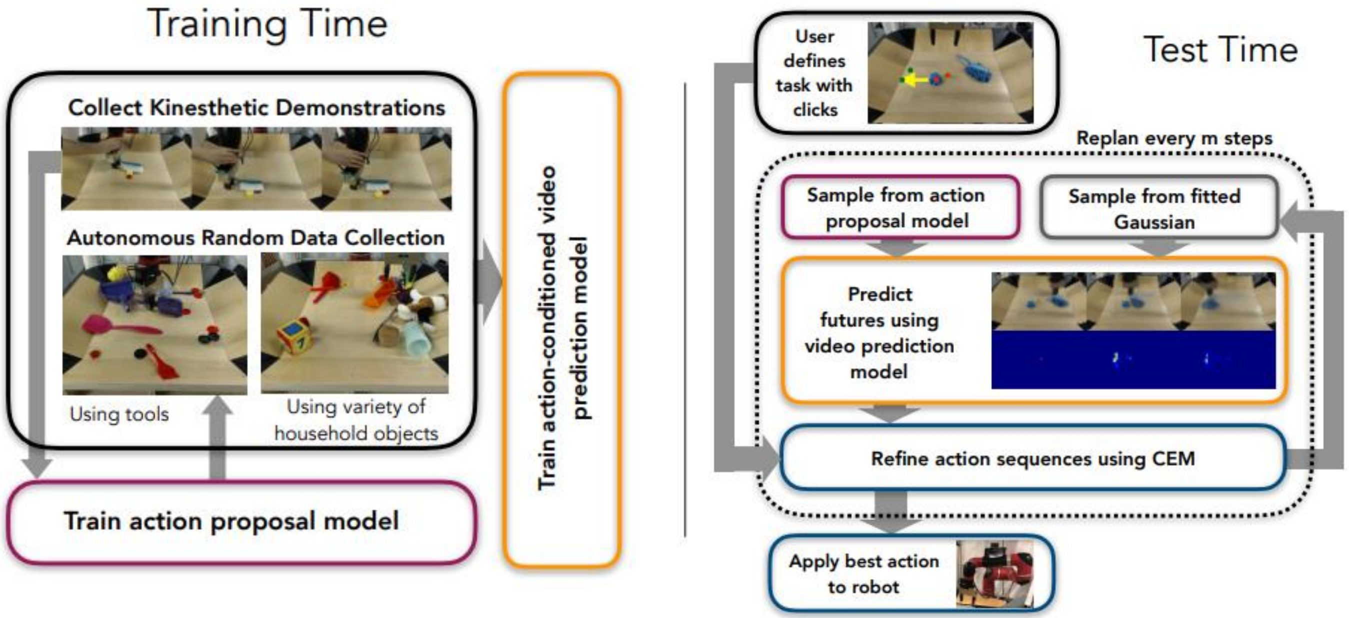



Ebert, Finn, Lee, Levine. **Self-Supervised Visual Planning with Temporal Skip Connections.**

# Model-based RL with improvised tool use

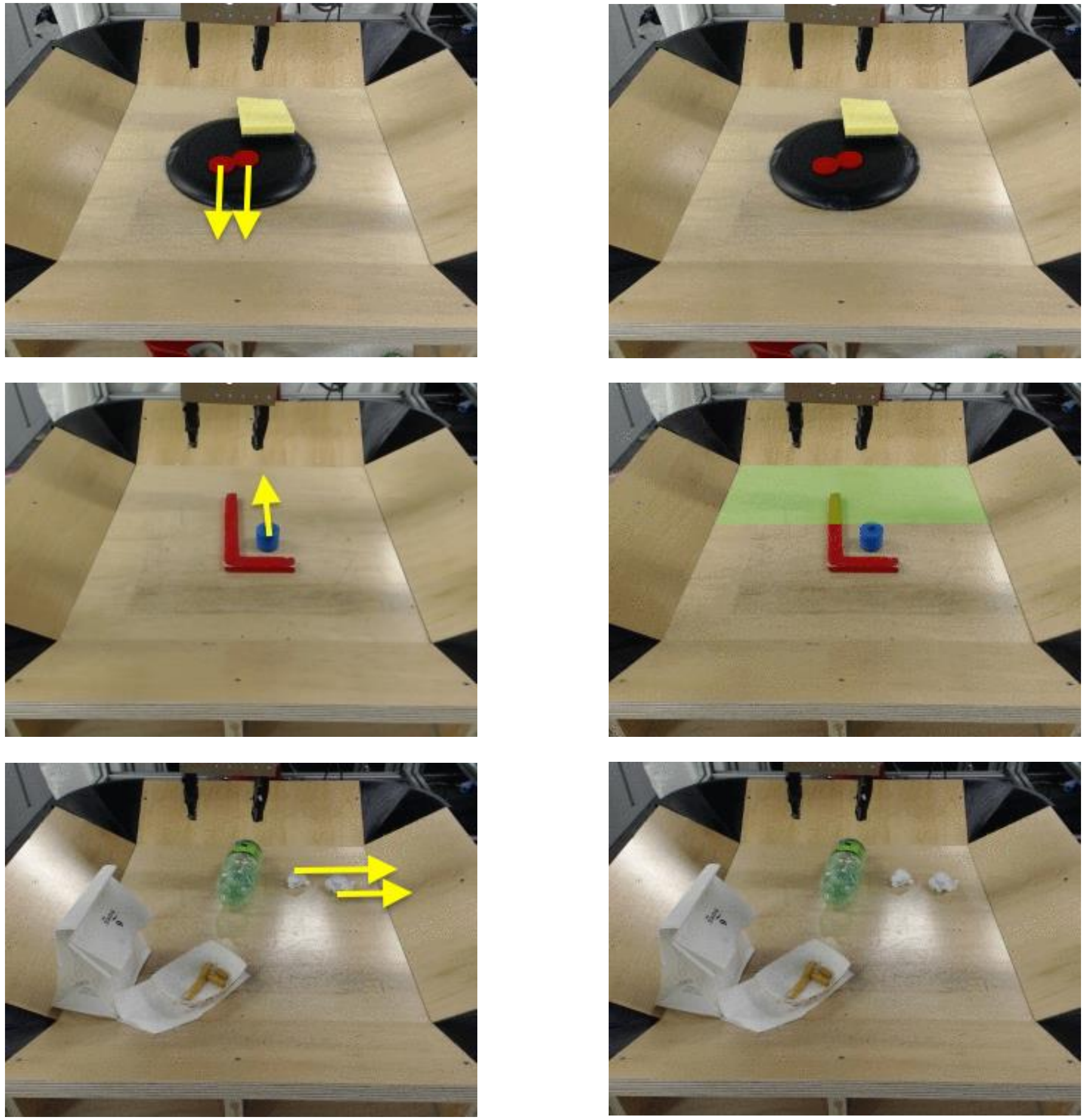






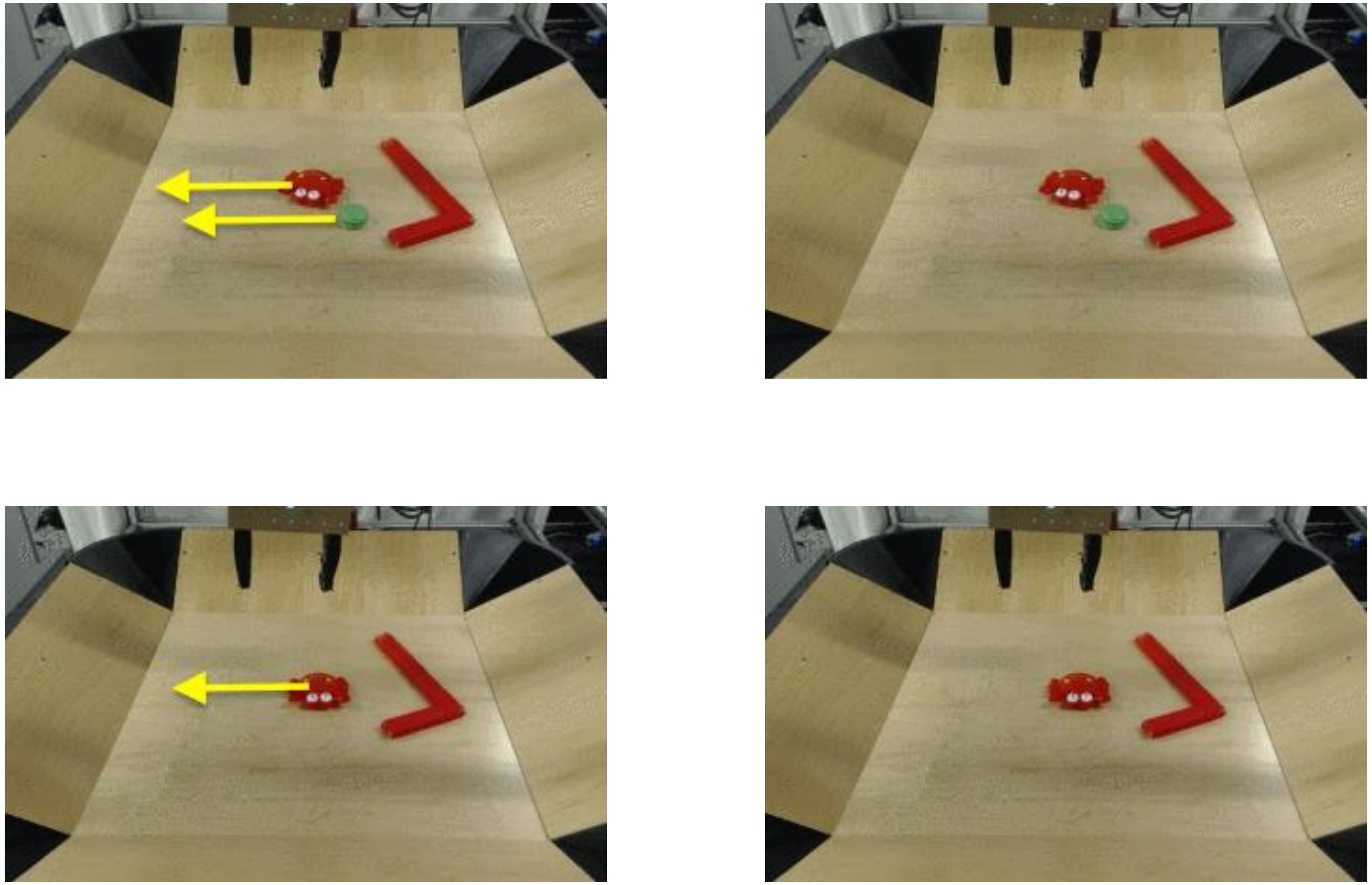

Xie, Ebert, Levine, Finn. **Improvisation through Physical Understanding: Using Novel Objects as Tools with Visual Foresight.** RSS '19

# Model-based RL with improvised tool use



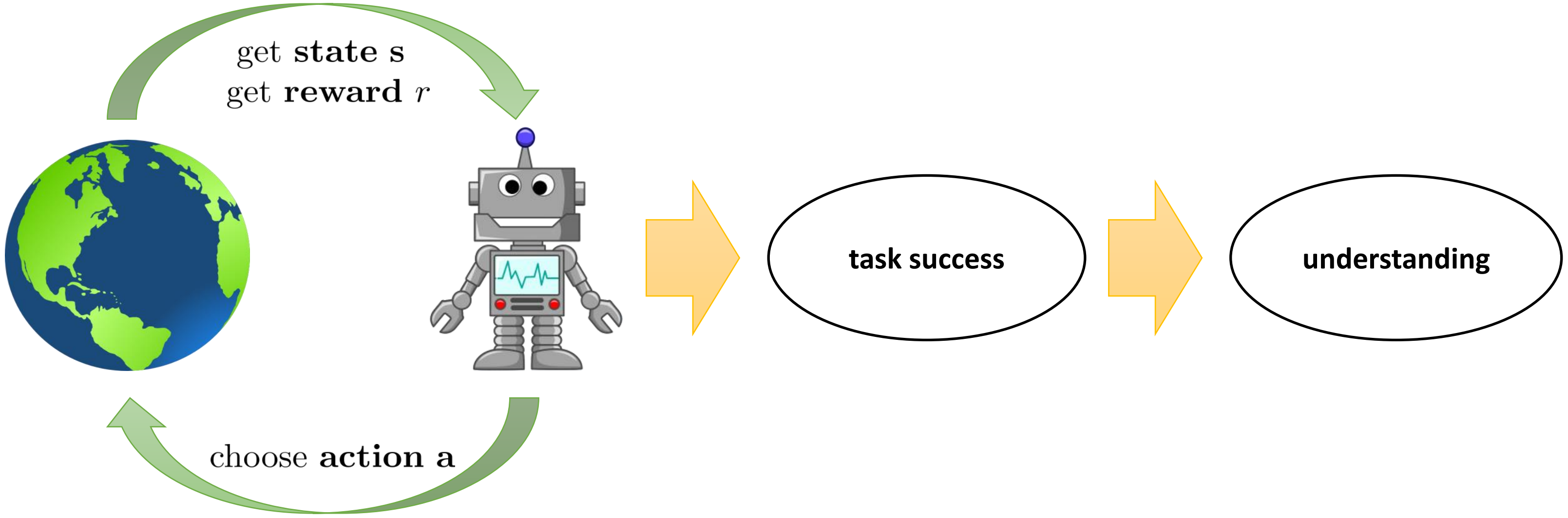
Xie, Ebert, Levine, Finn. **Improvisation through Physical Understanding: Using Novel Objects as Tools with Visual Foresight.** RSS ‘19

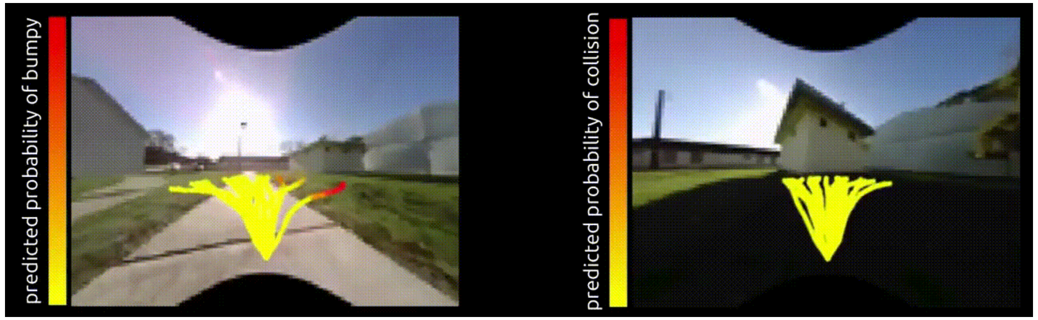
# Model-based RL with improvised tool use

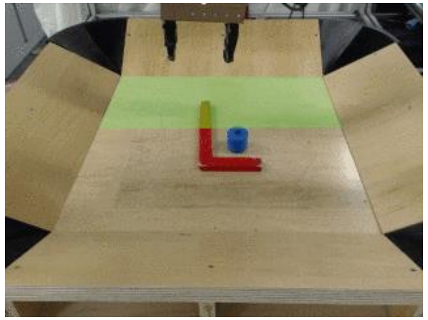


Xie, Ebert, Levine, Finn. **Improvisation through Physical Understanding: Using Novel Objects as Tools with Visual Foresight.** RSS ‘19

# An **embodied learning** recipe for scene understanding?



RAIL
Robotic AI & Learning Lab

website: http://rail.eecs.berkeley.edu
source code: http://rail.eecs.berkeley.edu/code.html